Novagen: A Combination of Eyesweb and an Elaboration-Network Representation for the Generation of Melodies under Gestural Control

Alan Marsden
Music Department, Lancaster University
Lancaster, LA1 4YW, UK
A.Marsden@lancaster.ac.uk

Abstract

A system which generates melodies influenced by the movements of a dancer is described. Underlying the melody-generation is a representation based on the theory of the early 20th-century German musicologist, Heinrich Schenker: a melody is derived from a simple background by layers of elaboration. Overall, the theory has similarities to a generative grammar. Generation of melodies is achieved by repeatedly applying elaborations to the background to achieve the desired number of notes. Elaborations are selected by a weighted random process which can take into account the pattern of elaborations used earlier in the melody. A number of parameters control this process, both by setting the relative weights for different elaborations and by controlling the number of notes generated, their distribution throughout the bar, and the degree of similarity of the generated pattern to previous sections of the melody. These parameters are adjusted via MIDI messages from an Eyesweb application which tracks a dancer via video to categorise the pose or movement observed into one of four categories, and to determine the degree of 'activity' in the movement. The result is real-time generation of a novel melodic stream which appears meaningfully related to the dancer's movements.

1 Introduction

In existing systems where gesture controls some form of musical output, it is not common for the music to be actually generated in response to the gestural input; more commonly either the gesture triggers, controls or otherwise modulates pre-composed music, so the metaphor for gestural control is conducting a musical ensemble, or the gesture triggers individual notes or sounds, and the metaphor is of playing a musical instrument. In this project, not only is the generation of a melody under gestural control, but the melody is intended to conform to the stylistic traits of common-practice tonal music, such as found in music from Bach to Brahms. This is achieved by basing the melody-generation on an established theory of the music of that period, i.e., the theory of Heinrich Schenker (1868-1935). The primary objective of the project is to test the underlying theoretical model as a formalisation of musical structure, a model which has potential applications in a number of musical spheres. If the objective had been principally creative or principally to develop a gestural interface, the path of research would have been different. Thus the approach taken towards the creation process does not have a high degree of artistic sophistication. On the other hand, in the area of gestural interface, a high degree of sophistication was found readily to hand in the form of Eyesweb (Camurri et al., 2000; www.eyesweb.org), which allowed the easy extraction of useful information from the movements of a dancer by the simple means of a video camera.

Overall, the project can be reported to have been successful, in that the essential concept is proven to be able to produce melodies whose characteristics recognisably change in relation to the movements of the dancer, and which remain stylistically 'correct': the music never sounds incoherent or 'wrong'. On the other hand, the melodies produced do not sound particularly musically appealing, and in particular they lack division into meaningful phrases. To date, the system has only been tested with a set of video

films of a dancer rather than with a live video feed, but in principle there is no reason why it should not work in this situation also, where there would also be the added advantage of feedback from the melody-generation to the dancer, allowing the possibility of creative interaction between the dancer and the system.

2 Underlying Musical Structure

2.1 Schenkerian Theory

A common theme of music theory from the eighteenth century has been that underlying the sequence of notes which forms the 'surface' of a melody is a less elaborate framework. The idea finds its fullest exploitation and culminating exposition in the work of Heinrich Schenker, whose seminar work *Der freie Satz* (1935). Computational implementations of the theory are found in the work of Kassler (1967), Frankel, Rosenschein & Smoliar (1976), and a number of more recent authors. Pursuing the common parallel between music and language, the theory has been compared to generative grammar, and a number of computational implementations of musical grammars have been reported also, some more closely related to Schenkerian theory (e.g., Baroni, 1983, and Baroni, Dalmonte & Jacoboni, 1992), and others of a very different nature (e.g., Kippen & Bel, 1992). The two ideas have come together also in the influential theory of Lerdahl and Jackendoff (1983), which has itself been subject to attempts at computer implementation (e.g., Baker 1989).

2.2 Representation System

The system used here was first reported in Marsden (2001) as a means of representing musical pattern. It differs fundamentally from the adaptation of Schenkerian theory by Lerdahl and Jackendoff in that elaborations are taken to apply to intervals between notes rather than to the individual notes of background and middleground structures. Indeed, the foundation of the system is a set of elaborations which could be expressed as rewrite rules whose left-hand sides are almost always a pair of notes and whose right-hand sides are three notes ('almost always' because in a few cases, such as suspensions, a context wider than just two notes must be taken into account). Thus each elaboration generates a new note. In some cases this note occurs between the two original notes. Thus an 'upper neighbour note' is a note one step higher than the original second note, and placed in time between the original first and second notes. (This is American/German terminology commonly used in Schenkerian writing; the traditional British terminology for the same thing is an 'upper auxiliary note'.) Some elaborations (referred to as 'accented elaborations') displace the first note so that it occurs later and put a new note in its place. Thus an appoggiatura (which can also be described as an accented upper neighbour note) is a note one step higher than the original first note which occurs at the original time of that note and is followed by a note of the same pitch as the original first note but placed in time between the original first and second notes. Manipulation of the representation is simplified by insisting that each elaboration produce no more than one new note, but this does mean that passing notes and other elaborations which produce more than one note can require a series of interdependent elaborations in the representation.

Figure 1, which is screenshot showing a fragment of melody generated by the system, demonstrates the principles of the system. The pair of notes on the top line are part of the underlying background. Each lower line shows a layer of elaboration on the way to the final melody shown in the bottom line. The boxes in between contain codes for the elaborations used in generating the melody.

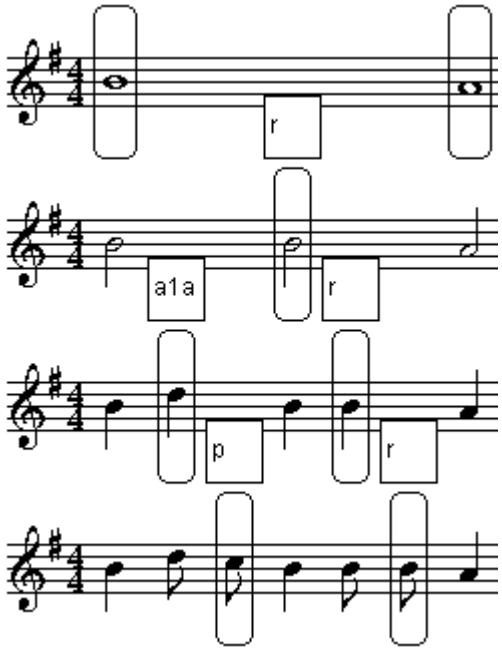

Figure 1. Example of generated melody

For each note in the melody, there is a prevailing key, harmony and metre. These influence the generation of new notes, so that, for example, what is meant by 'one step' depends on the pitch of the original note and the prevailing key: it might be a whole tone or a semitone. The placement of notes in time is determined by the metre and by a property of each elaboration which determines whether it is 'even', 'short-long' or 'long-short'. The last, for example, would produce dotted rhythms in a duple metre.

2.3 Benefits for Automatic Generation

This system of representation has three benefits for automatic generation. Firstly, just as a grammar ensures that any sequence generated by the rules of the grammar is 'grammatical', this representation system ensures that any melody generated by a process of elaboration from a simple background is 'musical', at least to the degree that the sequence lacks notes which sound 'wrong'.

Secondly, the derivation from an underlying framework ensures that the melody follows a logical harmonic pattern (e.g., a conclusion on the tonic can be guaranteed). Furthermore, it increases the likelihood of some sense of purposeful goal-directed motion in the melody, in contrast to the impression of aimlessness which can result from generation mechanisms such as stochastic processes which essentially append notes repeatedly to an existing sequence.

Thirdly, that the system explicitly represents musical pattern (as described in Marsden, 2001) allows explicit operations on musical patterns. Thus the pattern of an earlier part of the melody can be replicated by repeating its component elaborations in the new context. This will result in a different sequence of notes and a different sequence of intervals according to the interval and prevailing harmony of the new context: what is an interval of a third in one context might become an interval of a fourth in another, etc. (This kind of adjustment of intervals to preserve a pattern in a different context is a common characteristic of actual pieces of music.) Furthermore, a pattern need not be simply replicated, but it can be made more or less elaborate by the addition or deletion of elaborations. Thus the common musical procedures of variation can be implemented.

3 Generation Procedure

Generation begins from a given background, which may be specified by the user in advance. This background consists of a single note per bar, including a specification of the key and metre, and the harmony for each bar. The generation procedure repeatedly passes over this background, adding elaborations in real time, and playing the resulting melody. The selection of elaborations, including the decision of whether to elaborate or not, is made at random, using weights governed by a number of time-varying parameters

The first of these parameters is a target number of notes per bar. As elaboration proceeds to deeper levels, the target number of notes to be generated obviously decreases (since notes have already been generated at higher levels), so elaboration inevitably stops when the target number reaches zero. At higher levels, a second parameter of 'evenness' determines how evenly the targets are divided between 'left' and 'right'. Thus, for example, the target number for a whole bar might be eight notes. There is one note already given by the background framework at the beginning of the bar, and at the first layer of elaboration, a new note will be generated, let us say in the middle of the bar. Thus six more notes are to be generated at lower levels. If the 'evenness' is high, these will be distributed three in the first half of the bar and three in the second half. If, on the other hand, it is low, these might be distributed one in the first half of the bar and five in the second half. The 'evenness' parameter can vary with the level of elaboration in addition to time. Thus it is possible that at higher levels notes might be distributed evenly while at lower levels they are distributed unevenly.

Another set of parameters provides a likelihood profile of elaborations. This can also vary with level in addition to time. Thus a melody might be generated with a high likelihood of arpeggiations (leaps within the prevailing harmony) at higher levels but a high likelihood of passing notes at middle levels, and repeated notes most likely at the lowest levels.

A final set of parameters controls 'regularity', which refers to the degree to which elaborations in one part of the melody follow the pattern of those in a preceding part. Different parameters determine the degree of regularity with relation to different time intervals, so at one time the melody might copy the same pattern of elaborations from one bar to the

next while with a different setting of parameters the melody might copy the same pattern of elaborations four times within the same bar. This is implemented by determining a degree of likelihood that the chosen elaboration will be a copy of the elaboration found previously in the generated melody at each of the appropriate time intervals (quarter bar, half bar, full bar, etc.), and a residual degree of likelihood that the elaboration will be chosen at random according to the likelihood profile determined by the parameters described above.

This generation procedure has been implemented as a Java application, building on previous work to implement the representational system derived from Marsden (2001).

4 Gestural Control

4.1 Eyesweb

Eyesweb was chosen as the mechanism for capturing gesture for the following reasons. Firstly, it requires no special hardware and operates with standard video and computer equipment. Thus the movements of the dancer are not impeded in any way, and the system can be readily used in many circumstances. Secondly, it is a platform which is efficient and yet remarkably easy to use, following a paradigm of interconnecting processing units which is familiar to musicians who have used such software as MAX and Pd. Thirdly, it has built-in facilities for the extraction of information about the position and movement of a human figure. Finally, it has facilities for MIDI output, which makes real-time intercommunication with musical software very simple.

4.2 Gesture Capture

The Eyesweb application developed for this project assumes that the input is video of a dancer moving in an otherwise unchanging scene. For each frame, Eyesweb extracts the outline of the figure by subtraction from the background and can compute various pieces of information about the figure in space, such as the position of the limbs. For this application, however, the information used is the bounding rectangle, the area occupied by the figure within that rectangle, and its 'centre of gravity'. On the basis of this, two fundamental pieces of information are determined.

Firstly, the gait or posture of the dancer is placed into one of four categories. When 'walking' the figure occupies a large proportion of the bounding rectangle, whose sides are large in relation to its top and bottom, and the 'centre of gravity' moves at a moderate speed. When 'dancing', by contrast, the area occupied by the figure is smaller in relation to the bounding rectangle, because the dancer will often have arms or legs extended, and the rectangle may be more square, but the centre of gravity is once again moving. A third gait described as 'posed' is similar, but the centre of gravity moves little (the dancer might be moving the limbs, and so might not be posed in the strict sense of the word). Finally a 'crouching' gait is recognised by once again little movement of the centre of gravity but a large proportion of the bounding rectangle taken up by the figure. These four gaits were selected on the basis of the movements of the dancer observed in the video films used as experimental material in the course of this project.

The second piece of information extracted in the Eyesweb application is the degree of activity in the dance, related to the speed of movement of the dancer. This might be movement of the whole body from one place to another or movement of the limbs without moving the body. Thus the measure is based on the speed of the fastest-moving edge of the bounding rectangle. Movements of the whole body are likely to cause both the sides of the bounding rectangle to move in the same direction. Movements of the limbs are likely to cause just one side or the top of the rectangle to move, or perhaps both sides in opposite directions. Only movements towards or away from the camera, which are rare in isolation, cause no movement of the edges of the bounding rectangle. In order to compensate for the distance of the dancer from the camera, the speed of movement is scaled according to the size of the bounding rectangle.

The Eyesweb application therefore captures not so much individual gestures as global information about the characteristics of gestures at any one moment. This information is transmitted via MIDI control messages which indicate (1) the nature of the dancer's gait at that time, and (2) the speed of movement.

4.3 Control of Melody-Generation

The melody-generation application receives MIDI messages and responds to the control messages used

to transmit the gestural information by varying the parameters described above in section 3. The speed of movement of the dancer is related directly to the number of target notes: the faster the dancer moves, the more notes in the melody.

Gait, on the other hand is related to sets of other parameters so that generated melodies with different characteristics are associated with each gait, following metaphors of music and movement which appeared viable to the author. Thus, a crouching gait is associated with a high degree of regularity—so that the unchanging position of the dancer is reflected in the unchanging pattern of the music—and with small intervals—the compact shape of the dancer is related to the small degree of movement in pitch.

5 Conclusions

The overall musical results are, as predicted, melodies which sound 'musical' to the degree that they follow the kind of harmonic and intervalic patterns found in real music. The association with the movements of the dancer also seem credible in that changes in the dancer's pattern of movement are accompanied by meaningfully related changes in the melody. One does not have an impression of the dancer controlling the music, though, in part probably because there is not a sufficiently rapid and tight connection between the dancer's movements and events in the music. There is nothing, for example, which suggests that a particular note or set of notes has been generated specifically because of a particular gesture by the dancer—there is no sense of the dancer directly making things happen. The impression, rather, is of an independent musician who is responsive to the movements of the dancer and who varies the melody played according to the character of those movements.

The project demonstrates Eyesweb to be an effective and useable gesture-input system for this kind of application. A higher degree of control, along the lines mentioned in the previous paragraph, would have required a finer analysis of the position and movement of the dancer. Eyesweb does include such facilities, but, as described in section 4.2 above, for this project these facilities have not been used.

The major objective of the project was to test the underlying representation system as a basis for melody-generation, and in this it has been successful only to a degree. The melodies generated have some musical credibility, but the expectation that the generation of melodies by elaboration of a coherent framework would ensure a sense of goal-directedness has not been realised. The melodies also lack any sense of phrase (any sense of starting and stopping at certain points, or of division of the stream of notes into coherent melodic units). These two deficiencies are probably related. They might be rectified by a revision of the melody-generation process so that melodies are explicitly generated in phrases which in turn are made up of smaller melodic units, and so on. A disadvantage of this is that it would introduce a coarser level of granularity into the generation process and create difficulties for a design which aimed to see changes in the gestural input reflected quickly in the melodic output.

A second possible approach to overcoming the lack of goal-directedness and phrase structure is a more reflective melody-generation system. One pattern of notes can generally be generated as a result of a number of different patterns of elaboration, especially if different possible background frameworks are considered also. Thus while a melody might have been generated as a result of one pattern of elaborations, it might be perceived as a different pattern of elaborations. This is particularly important where a melody aims to show some degree of regularity by copying the pattern of elaborations used earlier: if the pattern of elaborations perceived by a listener is different from the pattern used by the generation process, then the regularity might not be perceived. Thus the melody-generation process needs to analyse its own output to consider how a sequence of notes might be perceived. There is a considerable quantity of research to be done here, most fruitfully probably in analysis of real melodies.

Overall, the general concept is proven effective, and the paradigm of a system of music-generation responding to a dancer's movements is shown to be aesthetically viable. One can imagine a more sophisticated system in future with which a dancer becomes familiar so that he or she can produce music and movement in a single unified yet improvised art work.

Acknowledgements

This research has been supported by a grant from the Arts and Humanities Research Board in 2001 under their scheme of Small Grants in the Creative and Performing Arts which paid for a visit to the University of Genoa. I am grateful to Antonio Camurri for his assistance and support, particularly in allowing me to work for a period in his laboratory at the University of Genoa, for guidance from his co-researchers in the use of Eyesweb, and for the use of video recordings of a dancer taken in Genoa.

References

Michael Baker. A Computational Approach to Modeling Musical Grouping. Contemporary Music Review, 4: 311-325, 1989.

Mario Baroni. The Concept of Musical Grammar. Music Analysis, 2: 175-208, 1983

Mario Baroni, Rossana Dalmonte & Carlo Jacoboni. Theory and Analysis of European Melody. In Alan Marsden & Anthony Pople (eds.) Computer Representations and Models in Music, London: Academic Press, 1992: 187-205.

A. Camurri, S. Hashimoto, M. Ricchetti, R. Trocca, K. Suzuki, G. Volpe. EyesWeb – Toward Gesture and Affect Recognition in Interactive Dance and Music Systems. Computer Music Journal, 24(1): 57-69, 2000.

R.E. Frankel, S.J. Rosenschein & S.W. Smoliar. A LISP-Based System for the Study of Schenkerian Analysis. Computers and the Humanities, 10: 21-32, 1976.

Michael Kassler. A Trinity of Essays. PhD thesis, Princeton University, 1967.

Jim Kippen & Bernard Bel. Modelling Music with Grammars: Formal Language Representation in the Bol Processor. In Alan Marsden & Anthony Pople (eds.) Computer Representations and Models in Music, London: Academic Press, 1992: 207-238.

Fred Lerdahl & Ray Jackendoff. A Generative Theory of Tonal Music. Cambridge, Mass.: MIT Press, 1983

Alan Marsden. Representing Melodic Patterns as Networks of Elaborations. Computers and the Humanities, 35:37-54, 2001.

Heinrich Schenker. *Der Freie Satz*. Vienna: Universal Edition, 1935. Published in English as *Free Composition*, translated and edited by Ernst Oster. New York: Longman, 1979.